基于 metaPRS 与 APOEε4 优化轻度认知障碍遗传风险统计建模策略的应用研究

李梓盟[1]，王荣[1]，陈帅[1]，赵彩丽[1]，王晓聪[3]，温雅璐[1,2*]，刘龙[1,2*]

基金项目：国家自然科学基金（81903418；82173632）

1.山西医科大学公共卫生学院卫生统计学教研室，030000，太原
2.重大疾病风险评估山西省重点实验室，030000，太原
3.澳大利亚莫纳什大学公共卫生与预防医学学院，3800，澳大利亚
*通信作者：温雅璐，教授，博士生导师；E-mail：wenyalu1031shanxi@126.com
刘龙，讲师，硕士生导师；E-mail：biostat-ll@sxmu.edu.cn

【摘要】 **背景** 轻度认知功能障碍（Mild cognitive impairment，MCI）是干预和延缓痴呆进展的重要阶段，既往研究发现 MCI 与遗传因素存在紧密关联，且载脂蛋白 E（Apolipoprotein E，APOE）ε4 是医学界公认的 MCI 重要风险等位基因。由于缺少 MCI 的全基因组关联研究（Genome-wide association study，GWAS）汇总数据，当前普遍以阿尔茨海默症（Alzheimer's disease，AD）的 GWAS 汇总数据作为 Base 数据集来计算 MCI 的多基因风险评分（Polygenic risk score，PRS），致使 MCI 的 PRS 遗传风险预测效果并不理想。**目的** 本研究以多基因遗传风险综合评分（Meta-polygenic risk score，metaPRS）与 APOEε4 作为重要预测因子，从广义线性模型与机器学习角度，探索并优化 MCI 的遗传风险统计建模策略。**方法** 计算 MCI 的 12 个亚表型 PRS，并利用弹性网状 Logistic 回归模型将其整合为MCI的metaPRS。利用年龄矫正的APOEε4效应量计算APOEε4加权总和（$SCORE_{APOE}$）。以metaPRS、$SCORE_{APOE}$ 及基本人口学信息（年龄、性别、受教育程度）构建不同的预测因子纳入策略，以 XGBoost，GBM，Logistic 回归及 Lasso 回归作为统计建模方法，采用 AUC 及 *F-measure* 评价 MCI 遗传风险统计建模的预测效果。**结果** metaPRS 与 $SCORE_{APOE}$ 对于 MCI 的遗传风险有较高的预测价值，纳入 metaPRS、$SCORE_{APOE}$ 及基本人口学信息（年龄，性别，受教育程度）后，各个统计建模方法的预测效果为：XGBoost（AUC=0.69，*F-measure*=0.88），GBM（AUC=0.76，*F-measure*=0.87），Logistic 回归（AUC=0.77，*F-measure*=0.89），Lasso 回归（AUC=0.76，*F−measure*=0.92）。**结论** 在样本量不高（小于 500）的情况下，以 metaPRS、$SCORE_{APOE}$ 与基本人口学信息为预测因子，以 Lasso 回归为统计建模方法的 MCI 遗传风险预测效果最好，为 MCI 等复杂疾病的遗传风险统计建模提供了新的思路与视角。

【关键词】 轻度认知障碍；多基因风险评分；MetaPRS；APOEε4；遗传风险预测；统计建模优化

Application of metaPRS and APOEε4 to optimize genetic risk prediction modeling strategy for mild cognitive impairment

LI Zimeng[1]，WANG Rong[1]，CHEN Shuai[1]，ZHAO Caili[1]，WANG Xiaocong[3]，WEN Yalu[1,2]，LIU Long[1,2*]*
1. Department of Epidemiology and Health Statistics，Shanxi Medical University，Taiyuan 030000，China
2. Shanxi Key Laboratory of Risk Assessment for Serious Diseases，Taiyuan 030000，China
3. School of Public Health and Preventive Medicine，Monash University，3800，Australia
**Corresponding author：WEN Yalu，Professor，Doctoral supervisor；E-mail：wenyalu1031shanxi@126.com*
LIU Long，Lecturer，Master supervisor；E-mail：biostat-ll@sxmu.edu.cn

【Abstract】 **Background** Mild cognitive impairment (MCI) is an important stage to intervene and delay the progression of dementia，and studies have shown that it is closely associated with genetic factors，among which Apolipoprotein E (APOE) ε4 is known to be an important risk allele of MCI in the medical community. Due to the lack of genome-wide association study (GWAS) summary data of MCI, existing studies calculate the polygenic risk score (PRS) of MCI based on GWAS summary data of Alzheimer's disease，which leads to the unsatisfactory effect of the existing statistical modeling of genetic risk of MCI. **Objective** In this study，meta-polygenic risk score (metaPRS) and APOEε4 were used as important predictors to explore and optimize the statistical modeling strategy of genetic risk in MCI from the perspective of generalized linear model and machine learning. **Methods** PRS for the 12 MCI-related traits were calculated and integrated into metaPRS for MCI by elastic-net logistic regression model. $SCORE_{APOE}$ is calculated by weighting the APOEε4 effect size with age correction. In this study，XGBoost，GBM，Logistic regression and Lasso regression were used as statistical modeling methods to verify the inclusion strategies of different predictors based on metaPRS，$SCORE_{APOE}$ and basic demographic information (age，gender，education level). AUC and *F-measure* were used to evaluate the predictive

effect of statistical modeling of genetic risk of MCI. **Results** For the genetic risk of MCI，metaPRS and $SCORE_{APOE}$ have high predictive value. After including metaPRS，$SCORE_{APOE}$ and basic demographic information (age，gender，education level)，the predictive effect of each statistical modeling method is as follows: XGBoost (AUC=0.69，*F-measure*=0.88)，GBM (AUC=0.76，*F-measure*=0.87)，logistic regression (AUC=0.77，*F-measure*=0.89)，and lasso regression (AUC=0.76，*F-measure*=0.92). **Conclusion** When the sample size is not high (less than 500)，the lasso regression model constructed by including metaPRS，$SCORE_{APOE}$ and basic demographic information (age，gender，education level) has the best effect on MCI genetic risk prediction，which provided a new idea and perspective for statistical modeling of genetic risk of MCI and other complex diseases.

【Key words】 Mild cognitive impairment；Polygenic risk score；MetaPRS；APOEε4；Genetic risk prediction；Statistical modeling optimization

轻度认知障碍（Mild cognitive impairment，MCI）是干预和延缓痴呆进展的重要阶段[1]。研究发现，MCI是遗传与环境因素共同作用的结果，且载脂蛋白E（Apolipoprotein E，APOE）ε4与MCI高度相关[2]。多基因风险评分（Polygenic risk score，PRS）是最常用的复杂疾病遗传风险预测方法之一，由于MCI特殊的疾病状态，尚无关于MCI的国际公开全基因组关联研究（Genome-wide association study，GWAS）汇总数据。目前普遍以阿尔茨海默症（Alzheimer's disease，AD）的GWAS汇总数据作为Base数据集用于MCI的PRS计算，导致MCI遗传风险的预测效果并不理想，关于MCI的遗传风险预测模型AUC普遍徘徊在0.58-0.68[3]。Abraham Gad [4]提出了多基因遗传风险综合评分（Meta-polygenic risk score，metaPRS），通过有效整合该疾病的多个亚表型PRS来进一步提高遗传风险的预测精度，且metaPRS已在缺血性脑卒中，抑郁症和冠心病等疾病得到很好的应用。此外，相关研究表明，基本人口学信息（年龄，性别，受教育程度）[5]和APOEε4加权总和（$SCORE_{APOE}$）[6]对MCI具有较高的预测价值，值得进一步探索。

MCI遗传风险统计建模方法主要包括广义线性模型（Generalized linear model，GLM）和机器学习（Machine learning，ML）两类。复杂疾病遗传风险预测统计建模通常需满足两个基本特性：一方面该模型可以处理非正态分布的表型，另一方面能够解决预测因子之间可能存在复杂函数关系问题。GLM中的Lasso回归是一种使用L1正则化的线性回归，与Logistic回归相比更具稀疏性，能够筛选重要的预测因子，且模型可解释性强。与GLM相比，ML中的XGBoost（Extreme gradient boosting）和GBM（Gradient boosting machine）则是通过训练多个弱监督模型后将其组合成为更稳健的强监督模型，更适用于捕捉变量间复杂的函数关系，但多数ML算法的内部结构并不透明，在可解释性方面劣于GLM。

本研究以metaPRS，$SCORE_{APOE}$与基本人口学信息作为MCI遗传风险统计建模的预测因子，特别是考虑到以上预测因子间可能存在的复杂函数关系及复杂的表型数据特征，从GLM和ML角度，以XGBoost，GBM，Logistic回归及Lasso回归作为统计建模方法，探索并优化MCI遗传风险统计建模策略，为MCI等复杂疾病的高危人群识别，早期预防与干预，及精准医学研究提供新的视角和科学依据。

1 材料和方法

1.1 数据来源与质量控制

1.1.1 数据来源 关于MCI遗传风险预测研究所需的基因组学数据，来自于英国生物数据库（United Kingdom Biobank，UKB）与阿尔茨海默症神经成像计划（Alzheimer's Disease Neuroimaging Initiative，ADNI）。UKB是一个大型前瞻性队列研究及生物医学数据库，主要收集了认知功能测试，血压，身体测量数据，血液检查数据，基因测序数据，全身影像数据（例如：脑部MRI与心脏MRI）和随访数据等多方面的数据。ADNI是一项大规模的队列研究，主要收集了受试者的人口统计学变量（例如：年龄，性别，受教育程度），脑部影像学数据，生物学标志物和基因测序数据。

本研究主要集中于脑结构成像表型，不仅准确选取了四种主要的大脑组织结构：白质体积（White matter），灰质体积（Grey matter），脑脊液体积（Cerebrospinal fluid，CSF）和总脑体积（Total brain）；还纳入了既往研究表明与MCI相关的脑结构成像表型：白质高信号体积（White matter hyperintensities，WMH），苍白球体积（Pallidum），尾状核体积（Caudate），海马体体积（Hippocampus），杏仁核体积（Amygdala），伏隔核体积（Accumbens），壳核体积（Putamen）和丘脑体积（Thalamus）[7, 8]。

1.1.2 质量控制 由于UKB数据库均为白人，为控制人口结构可能带来的混杂，经过主成分分析（见附录材料1-3，http://cstr.cn/31253.11.sciencedb.j00150.00009），本研究只保留了ADNI数据库中的非西班牙裔白人，使得本研究使用的ADNI与UKB数据库在人口结构上相似。采用PLINK 1.9去除个体缺失率大于10%的人群，去除基因型缺失率大于10%的SNP_S。数据质量控制后，根据物理位置提取UKB与ADNI数据库共同的SNP_S。最终UKB数据库得到488371个个体，694020个SNP_S，以此对各亚表型进行GWAS研究获得所需的GWAS汇总数据；ADNI数据库

得到 325 个个体，694020 个 SNPs。

1.2　方法　本研究从研究设计上分为三个阶段，见图 1。第一个阶段：在 ADNI 数据集中，计算 MCI 的 12 个亚表型 PRS。第二个阶段：基于弹性网状 Logistic 回归模型整合 12 个亚表型 PRS，并计算 MCI 的 metaPRS。第三个阶段：通过 10 折交叉验证对不同预测因子纳入策略及不同预测方法性能进行验证。

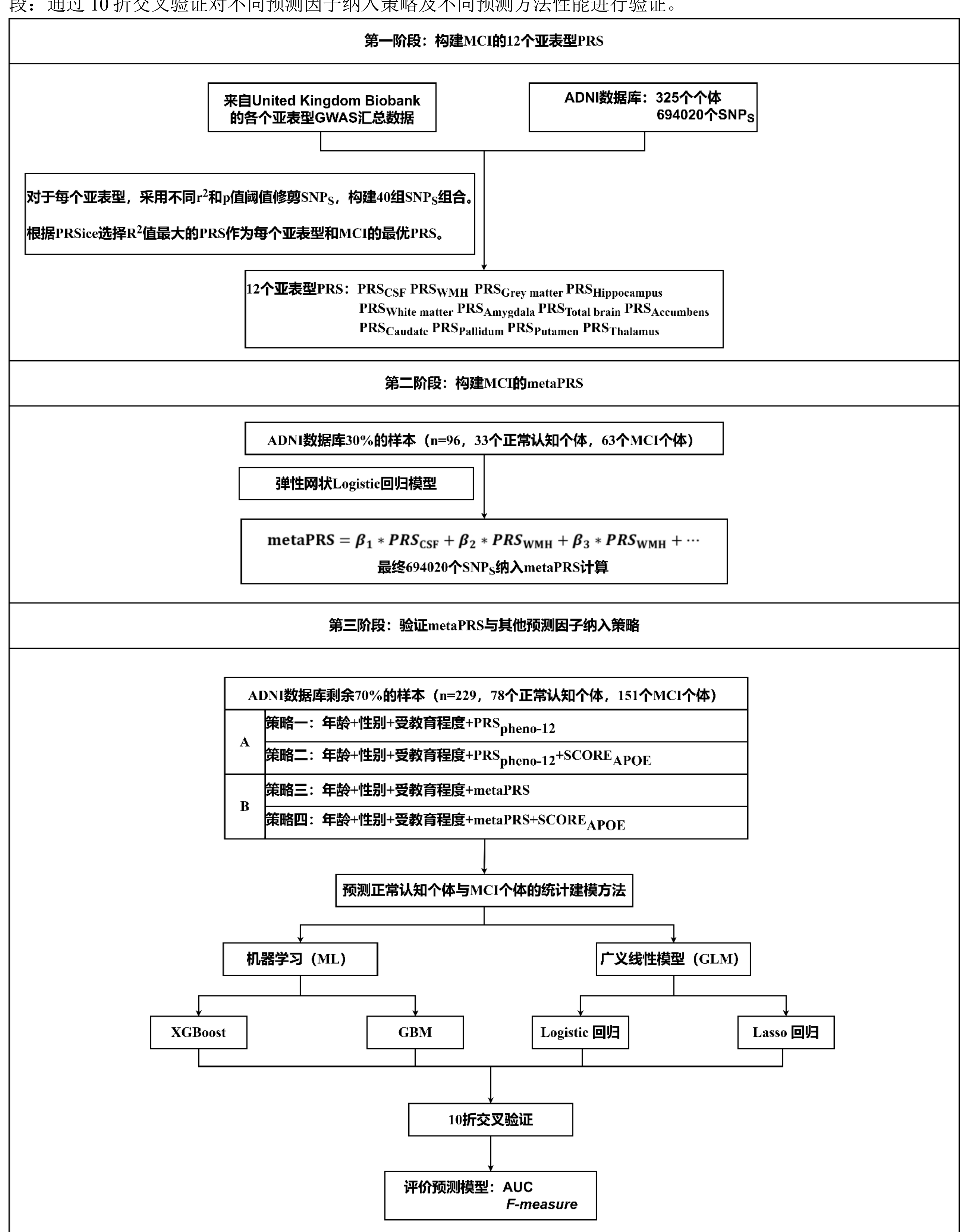


图 1　研究设计

Figure 1　Study design

1.2.1 全基因组关联研究 GWAS 将单核苷酸多态性（Single nucleotide polymorphisms，SNPs）位点与性状进行群体水平的统计学分析，识别和描述 SNPs 与疾病进展或疾病结局之间的关联[9]。GWAS 的研究结果主要通过 Quantile-Quantle（Q-Q）图和曼哈顿图进行可视化，曼哈顿图表示 SNPs 的显著性水平，Q-Q 图表示在 SNPs 水平上检验统计量的期望和分布之间的关系，λ-统计量评估是否有必要纳入主成分控制群体分层[9]。

1.2.2 metaPRS 的构建 （1）使用 PRS 的经典构建方法（Clumping and threshold，C+T）计算各个亚表型 PRS，PRS 为每个 SNPs 的风险等位基因的个数乘以其各自的效应量，构建公式为$PRS_i=\sum_{j=1}^{m}\beta_j X_{ij}$，$i$是第$i$个个体，$j$是第$j$个 SNP，$\beta$是 GWAS 汇总数据的效应值，$X_{ij}$是第$i$个个体第$j$个 SNP 风险等位基因的个数。

（2）在 ADNI 数据库（n=325）中随机抽出 30%的个体，采用弹性网状 Logistic 回归模型整合 12 个亚表型 PRS，在最终模型中获得每个亚表型 PRS 的系数$(\beta_1,\cdots,\beta_k)$作为权重[4, 10]以构建 metaPRS 的预测模型。

（3）利用$\beta_{snp_i}=\frac{\beta_1}{\sigma_1}\alpha_{j1}+\cdots+\frac{\beta_k}{\sigma_k}\alpha_{jk}$将亚表型 PRS 水平的权重转换为 SNPs 水平的权重，其中，$\sigma_1,\cdots,\sigma_k$是训练集中每个亚表型 PRS 的标准差，$\alpha_{j1},\cdots,\alpha_{jk}$是第$i$个 SNP 的等位基因对应于每个亚表型的效应值，如果第k个评分中未包含某个 SNP，则该 SNP 的效应值大小α_{jk}设为 0。

（4）根据公式$\text{metaPRS}=\sum\beta_{snp_i}\times N_i$计算 metaPRS，其中，$\beta_{snp_i}$是第$i$个 SNP 的效应值，$N_i$是个体所携带第$i$个 SNP 的效应等位基因数目。

1.2.3 预测因子纳入策略 本研究的预测因子纳入策略基于基本人口学信息和遗传信息进行构建，由于在 APOEε4 的连锁不平衡区域中 rs429358 是最显著的位点，所以选择 rs429358 代表 APOEε4[11]。且 APOEε4 的等位基因频率随着年龄的变化而变化[12]，所以本研究选择通过$\beta_{APOE\varepsilon4}=\ln OR$计算以年龄矫正的 APOEε4 效应量（个体年龄≤60，$\beta_{APOE\varepsilon4}=0.542$；60<个体年龄≤70，$\beta_{APOE\varepsilon4}=0.419$；70<个体年龄≤80，$\beta_{APOE\varepsilon4}=0.577$；80<个体年龄，$\beta_{APOE\varepsilon4}=0.425$ [13]），并单独计算 APOEε4 的加权总和[6]，其公式为$SCORE_{APOE-i}=\beta X_i$，其中，i是第i个个体，β是 APOEε4 的效应量，X_i是第i个个体 rs429358 风险等位基因的个数。本研究的预测因子纳入策略见表 1。

表 1 MCI 遗传风险统计建模预测因子纳入策略

Table 1 Description of predictor inclusion strategies for statistical modeling of genetic risk for MCI

策略序号	预测因子纳入策略
策略一	年龄+性别+受教育程度+ PRS_{pheno_12}
策略二	年龄+性别+受教育程度+ PRS_{pheno_12}+ $SCORE_{APOE}$
策略三	年龄+性别+受教育程度+ metaPRS
策略四	年龄+性别+受教育程度+ metaPRS+ $SCORE_{APOE}$

注：PRS_{pheno_12}，通过 UKB 的 GWAS 汇总数据构建的 12 个亚表型 PRS；metaPRS，整合 MCI 的 12 个亚表型 PRS 得到的 metaPRS；$SCORE_{APOE}$，APOEε4 的加权总和。

1.2.4 统计建模方法 （1）XGBoost（Extreme gradient boosting）算法是陈天奇博士提出的基于集成学习的 ML 算法[14]。XGBoost 的基本思想是利用函数的二阶导数信息来训练树模型，并把树模型复杂度作为正则化项加到目标函数中，使学习到的模型泛化能力更高。其目标函数为:

$$L(\varphi)=\sum_i l(\hat{y}_i,\ y_i)+\sum_k \Omega(f_k)$$

损失函数为$l(\hat{y}_i,\ y_i)$，正则化项为$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2$,其中，T代表叶子节点的个数，$\omega$表示叶子节点的分数。正则化项表示树的复杂度的函数，值越小，则复杂度越低，泛化能力越强。

（2）GBM（Gradient boosting machine）是 ML 中常用算法，该算法由大量简单的决策树集合而成，利用迭代多棵决策树不断学习残差，减小损失函数的值来调整模型，具有较高的可解释性[15]。GBM 在没有先验数据结构的情况下，可以对表型及其预测因子之间的关系建模，被认为是一种泛化能力较强的算法。GBM 可以表示为一组加性回归模型：

$$y^*=1\mu+\sum_{m=1}^{M}\vartheta h_m(y^*;X)+e$$

其中，y^*是表型，X 是预测因子，e是残差，ϑ用于控制每次迭代时从残差中减去的方差，从而在模型数量和预测因子相关性之间进行权衡。实际上，较小的ϑ需要组合更多的模型，以在训练集中达到相同的错误率，且会在验证集上产生更好的预测性能。

（3）Logistic 回归是预测结局变量为二分类变量时最为常用的统计模型，其通用形式如下：

$$\mathrm{Logit(P)} = \mathrm{Log}\left(\frac{P}{1-P}\right) = a + b_1x_1 + \cdots + b_mx_m$$

其中，$x_1,\cdots,x_m$为预测因子，$b_1,\cdots b_m$为m个预测因子的回归系数。Logistic 回归表达式经过简单变换，可得预测事件的概率P，表达式为$P = \frac{\exp(a+b_1x_1+\cdots+b_mx_m)}{1+\exp(a+b_1x_1+\cdots+b_mx_m)}$。

（4）Lasso 回归由 Tibshirani 于 1997 年提出[16]，旨在构建性能最佳的惩罚线性模型。在 Lasso 回归中较大的惩罚会导致一些预测因子的回归系数接近零，回归系数变为零的预测因子会从模型中删除。Lasso 回归具有较强的稀疏回归系数向量的能力，为模型选择有用的特征，具有更高的模型性能。

1.3 统计学分析 所有统计学分析均通过 R 软件（版本 4.1.0）完成。XGBoost，GBM，Logistic 回归和 Lasso 回归分别采用 XGBoost 包，gbm 包，stats 包和 glmnet 包。所有预测模型采用 10 折交叉验证方法验证预测性能，评价指标采用 F1 分数（*F-measure*）与 AUC。*F-measure* 是常用于评价二分类模型的信度指标，其数值越大，表示模型对于精确率和召回率的平衡效果越好且分类模型信度越高。

2 结果

2.1 研究对象的基本信息 病例组平均年龄（70.66±7.00）岁，对照组平均年龄（74.26±5.69）岁，病例组 APOEε4 等位基因频率为 45.79%，对照组 APOEε4 等位基因频率为 27.93%，见表 2。

表 2 ADNI 325 个受试者的一般情况

Table 2 General condition of 325 participants in ADNI

	正常认知个体（N=111）	**MCI 个体（N=214）**
年龄（岁）	74.26±5.69	70.66±7.00
性别（男/女）	59/52	114/100
受教育时间（年）	16.42±2.54	16.20±2.66
APOEε4 等位基因	31（27.93%）	98（45.79%）

2.2 全基因组关联研究 本研究计算了 12 个亚表型的 λ-统计量且其都接近于 1，这说明群体分层得到了适当的调整，见图 2。Amygdala，Caudate，CSF，Pallidum，Putamen 及 WMH 表型存在达到 Bonferroni 显著水平 $p<5\times10^{-8}$（第一条水平线）的 SNPs，这些 SNPs 位点所在的基因是 AD 的候选基因[17]。Accumbens，Grey matter，Hippocampus，Thalamus，Total brain 及 White matter 表型，在 $p<5\times10^{-6}$（第二条水平线）的阈值水平上存在许多显著相关的 SNPs。

本研究选用$p<5\times10^{-8}$这一阈值是可靠的，目前尚没有在该阈值下的SNPs被证明为假阳性[18]。Reed[9]在$p<5\times10^{-6}$的关联阈值下发现一些显著的SNPs，相较于$p<5\times10^{-8}$，$p<5\times10^{-6}$是不太严格的关联阈值，$p<5\times10^{-6}$的这些SNPs需要进一步验证，此方法类似于 Edmondson 的研究[19]。所以我们基于既往研究选择了 Bonferroni 显著水平（$p<5\times10^{-8}$）和 Bonferroni 阈值水平（$p<5\times10^{-6}$）用于判断多个亚表型 GWAS 汇总数据是否有研究价值的 SNPs。

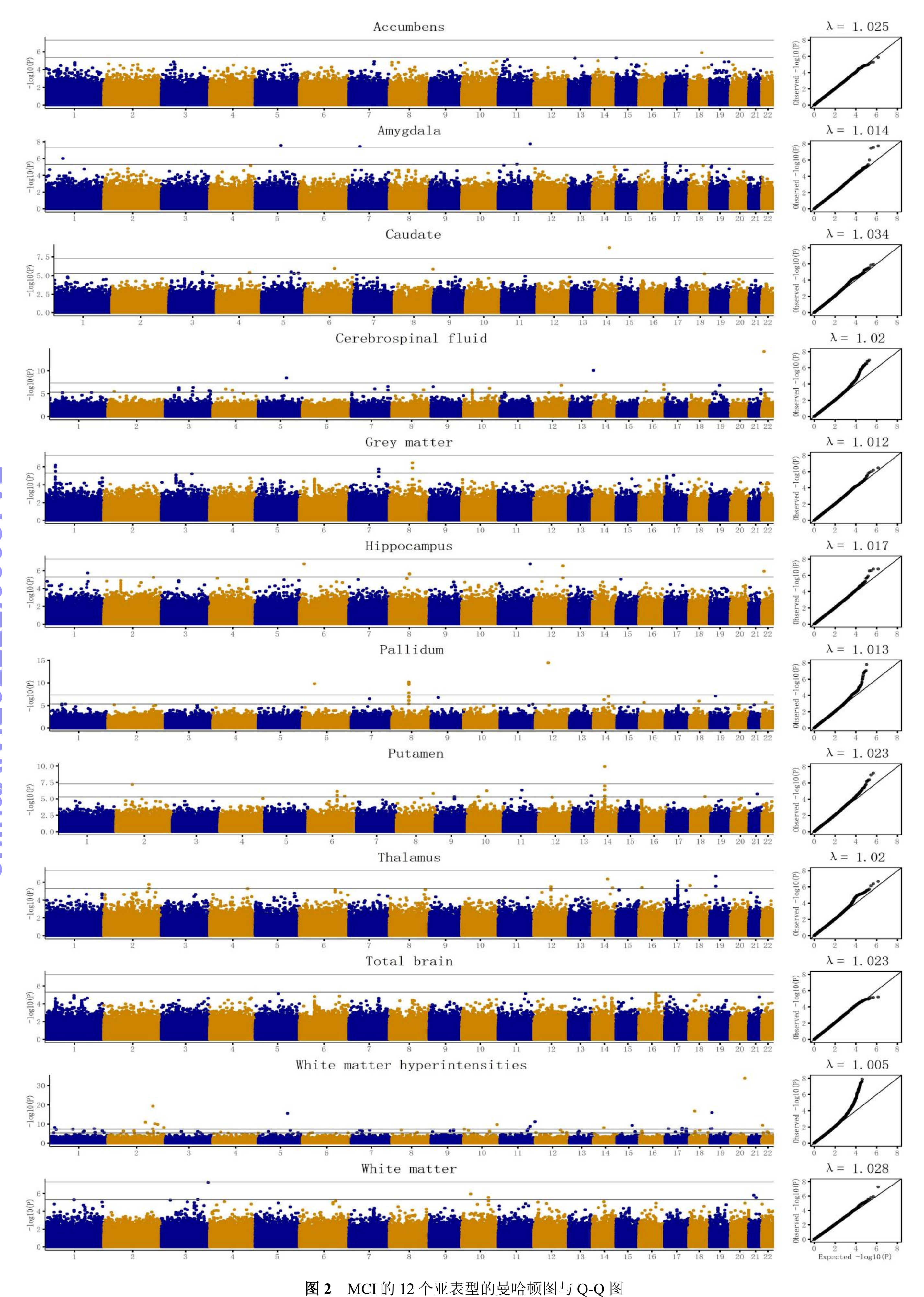


图 2　MCI 的 12 个亚表型的曼哈顿图与 Q-Q 图

Figure 2　Manhattan plots and Q-Q plots for 12 MCI-related traits

2.3 metaPRS 的构建 计算各预测因子之间的 *Pearson* 相关系数，如图 3 所示，各预测因子之间存在不同程度的相关性，其中，$PRS_{Hippocampus}$ 和 metaPRS（r=-0.6）、PRS_{WMH} 和 metaPRS（r=0.5）、$PRS_{Pallidum}$ 和 metaPRS（r=-0.5）、PRS_{CSF} 和 $PRS_{Accumbens}$（r=-0.4）、PRS_{CSF} 和 $PRS_{Total\ brain}$（r=-0.4）、$PRS_{Total\ brain}$ 和 $PRS_{Grey\ mattter}$（r=-0.4）以及 $PRS_{Accumbens}$ 和 $PRS_{Thalamus}$（r=0.4）。

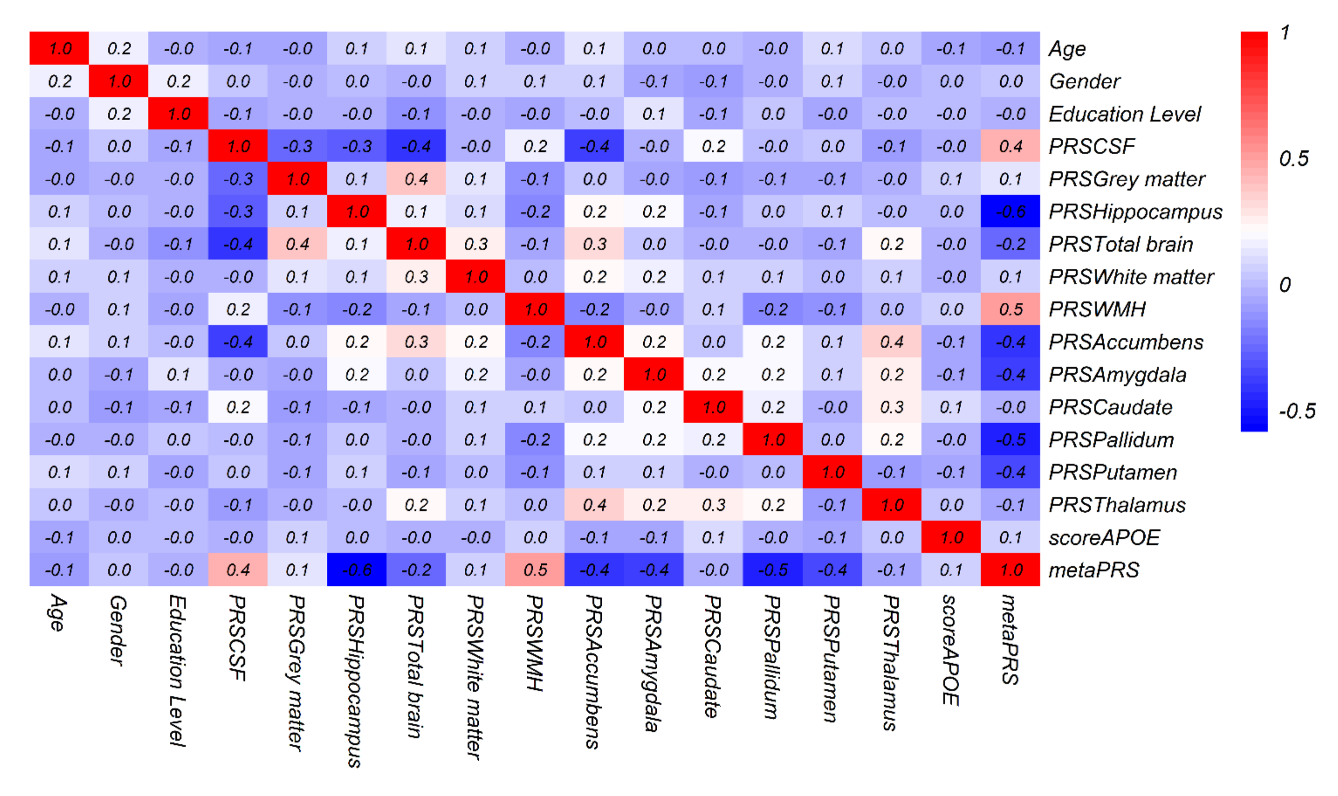


图 3 不同预测因子之间的 *Pearson* 相关系数

Figure 3 *Pearson* correlation coefficient of different predictors

2.4 不同预测因子纳入策略的验证 A 组（策略一 VS 策略二）和 B 组（策略三 VS 策略四），分别比较策略一和策略二以及策略三和策略四，加入 $SCORE_{APOE}$ 策略的趋势明显高于未加入 $SCORE_{APOE}$ 策略，说明 APOEε4 预测 MCI 的作用得到了验证。C 组（策略二 VS 策略四）通过比较策略二和策略四，在 4 种统计建模方法上策略四的趋势高于策略二，即基于 metaPRS 优化的预测因子纳入策略优于基于 12 种亚表型的 PRS 的预测因子纳入策略，如图 4 所示。

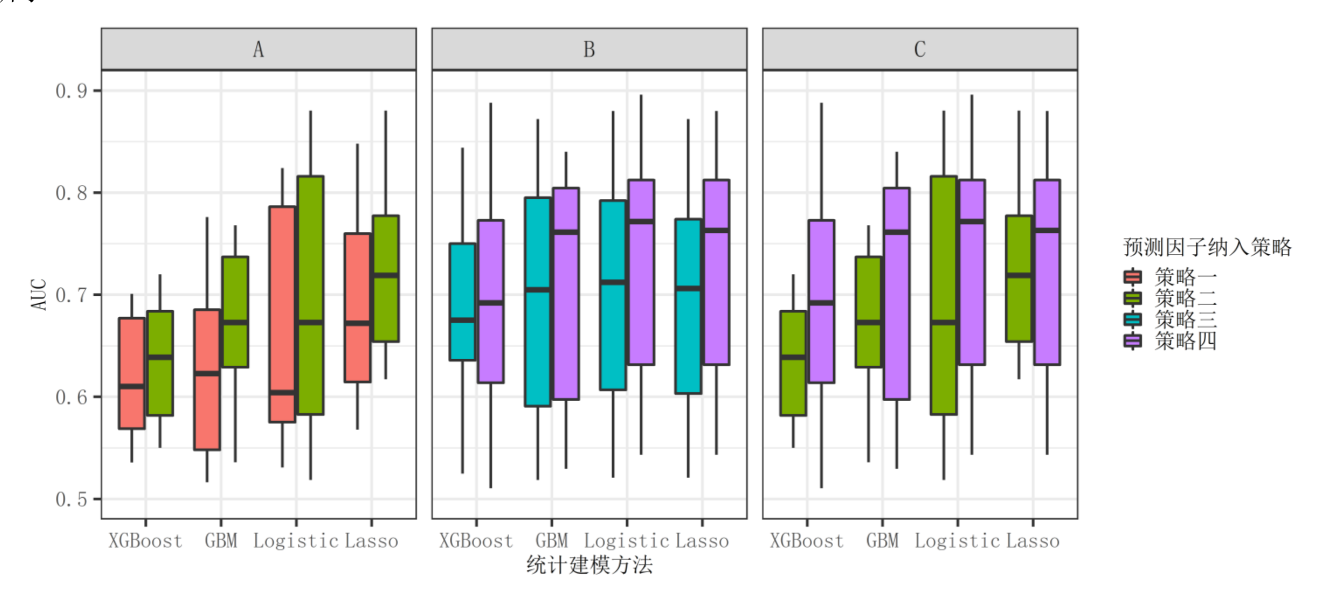


图 4 不同预测因子纳入策略的比较

Figure 4 Comparison of inclusion strategies for different predictors

2.5 统计建模效果的评价 总体来说，Lasso 回归的预测性能高于其他 3 种统计建模方法。A 组中，不同预测因子纳入策略下 Lasso 回归的 *F-measure* 高于其他 3 种统计建模方法；在策略四（metaPRS 和 $SCORE_{APOE}$）中，不同统计建模方法的 *F-measure* 分别为：XGBoost（*F-measure*=0.88），GBM（*F-measure*=0.87），Logistic 回归（*F-*

measure=0.89），Lasso 回归（*F-measure*=0.92）。B 组中，在策略四上不同统计建模方法的 AUC 离散程度大体一致，其中位数分别为：XGBoost（AUC=0.69），GBM（AUC=0.76），Logistic 回归（AUC=0.77），Lasso 回归（AUC=0.76）。

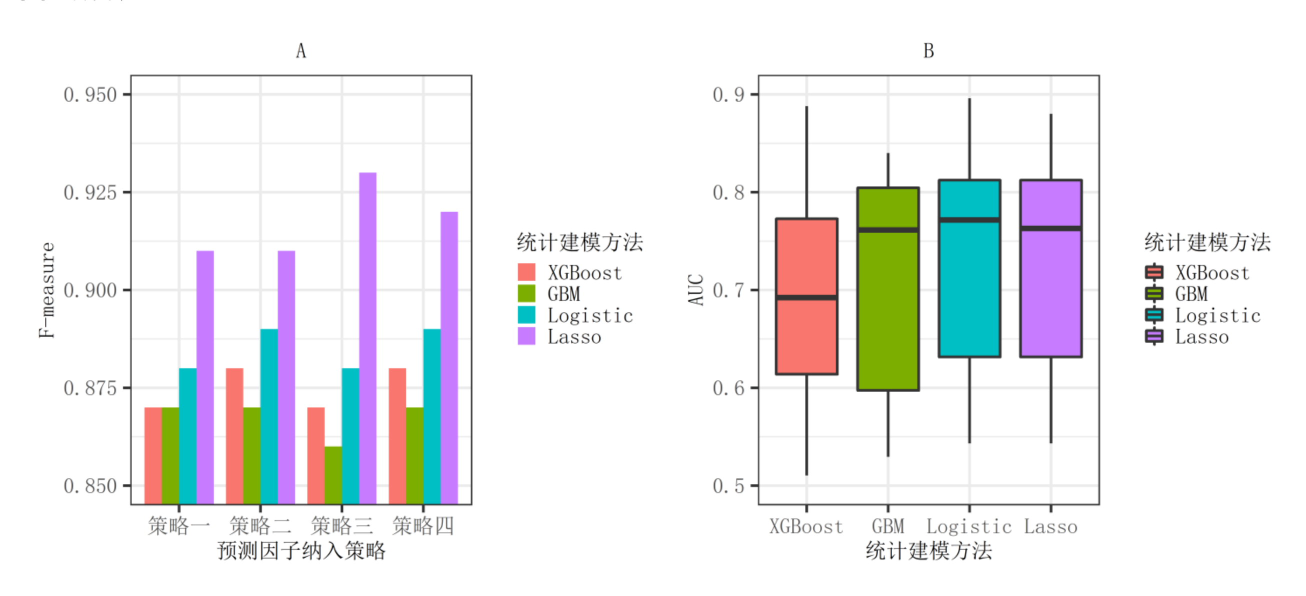


图 5 不同统计建模方法的 *F-measure* 与 AUC 比较

Figure 5 Comparison of *F-measure* and AUC of different statistical modeling methods

3 讨论

本研究以 12 个亚表型的 PRS，metaPRS，$SCORE_{APOE}$ 及基本人口学信息作为 MCI 统计建模的预测因子，以 XGBoost，GBM，Logistic 回归及 Lasso 回归作为 MCI 统计建模的方法，探索并构建了适用于 MCI 遗传风险预测的统计建模策略。特别是，研究发现，metaPRS 与 $SCORE_{APOE}$ 对于 MCI 的遗传风险具有较高预测价值，且在样本量不高（小于 500）的情况下，Lasso 回归是 MCI 遗传风险统计建模比较理想的方法。

研究发现，APOEε4 效应量进行年龄矫正后加权算分并作为预测因子纳入预测模型会显著提高 MCI 的预测分类效果，这说明 $SCORE_{APOE}$ 对预测 MCI 具有重要意义。已有研究表明在人群中 APOEε4 的等位基因频率会随着年龄的增长而下降且 APOEε4 效应量受年龄影响[12]，本研究再次验证利用经过年龄矫正的 APOEε4 效应量并加权算分作为独立预测因子纳入预测模型的合理性与科学性。本研究还发现，基于 metaPRS 的预测因子纳入策略优于基于 12 个亚表型 PRS 的预测因子纳入策略及既往对于 MCI 的预测策略，且基于 metaPRS 和 $SCORE_{APOE}$ 的预测因子纳入策略优于其他 3 种预测因子纳入策略（见图 4）。既往对于 MCI 的预测都是通过 AD 的 GWAS 汇总数据构建的 PRS，使用 AUC 评估预测效果在 0.58-0.68[3]。这是因为既往的 GWAS 汇总数据是关于 AD 的二分类变量，而本研究是选取与 MCI 相关的 12 个脑成像表型，合理整合有相关性的 12 个亚表型 PRS 构建 metaPRS，并选用了 XGBoost，GBM，Logistic 回归及 Lasso 回归进行分析比较，最终得到性能高的模型。因此，在未来 MCI 遗传风险预测的研究中，我们可以更多的关注相关预测因子的挖掘及整合预测因子方法的开发。虽然 MCI 的预测模型尚未达到临床诊断所需的水平，但与之前的研究相比，我们的分析取得了积极进展。

本研究综合 *F-measure* 与 AUC 两个评价指标，Lasso 回归的预测效果最好。一方面，在策略一（MCI 的 12 个亚表型）和策略二（MCI 的 12 个亚表型和 $SCORE_{APOE}$）中，Lasso 回归优于其他 3 种统计建模方法，主要是 Lasso 回归具有更强的稀疏回归系数向量的能力，惩罚线性回归更适用于有相关性的多个亚表型构建遗传风险预测模型。另一方面，在策略三（metaPRS）和策略四（metaPRS 和 $SCORE_{APOE}$）中，XGBoost 劣于其他 3 种统计建模方法，原因可能是本研究的样本量较小，XGBoost 相比于 Lasso 回归需要更大的样本量才能体现其性能优势。Christodoulou 等人做了一项综述研究[20]，汇集了 75 项研究的数据，其样本量中位数为 1250（样本量范围为 72-3994872），最终发现相比于 Logistic 回归，ML 在预测结果上没有明显优势。相关研究也表明[21]，在多种 ML 方法（朴素贝叶斯，XGBoost，支持向量机等）中，XGBoost 的性能最佳，但其预测效果非常依赖于样本量大小，在样本量小于 500 的情况下，与其他 ML 方法相比没有明显优势。

由于本研究训练集样本量不够大，这可能会影响研究结果的泛化能力，且本研究的基因组学数据是来自 UKB 和 ADNI 两个数据库交叉合并的共同物理位置 SNPs，可能会遗失与 MCI 相关的遗传信息。因此，建议未来基因测序数据考虑一些罕见变异的测量。此外，本研究仅采用了 4 种统计建模方法，未来将进一步探索其他可能提高 MCI 遗传风险预测精度的方法，并考虑构建全新的统计模型。

综上，以 metaPRS、$SCORE_{APOE}$ 与基本人口学信息（年龄，性别和受教育程度）作为预测因子，以 Lasso 回归

作为 MCI 遗传风险统计建模方法的统计建模策略取得了较理想的预测效果，有助于为 MCI 精准医疗及早期干预提供科学依据，具有一定的临床应用价值。必要情况下，将 MCI 的遗传风险预测作为健康体检项目或者相关门诊的常规筛查，可以在很大程度上提高 MCI 的检出率，进而实现 MCI 的早期干预，有效降低家庭及社会的疾病负担。

作者贡献：李梓盟负责提出研究选题方向、可行性分析、对结果解释分析，进行论文撰写、修订；王荣、陈帅，赵彩丽负责文献/资料收集、翻译与整理；王晓聪负责搜集数据；温雅璐，刘龙负责核心督导，对文章整体负责。所有作者确认了论文的最终稿。

本文无利益冲突。

参考文献

[1] ANDERSON N D. State of the science on mild cognitive impairment (MCI) [J]. CNS spectrums, 2019, 24(1): 78-87.

[2] LUO Y, TAN L, THERRIAULT J, et al. The Role of Apolipoprotein E ε4 in Early and Late Mild Cognitive Impairment [J]. European Neurology, 2021, 84(6): 472-80.

[3] LEONENKO G, SHOAI M, BELLOU E, et al. Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition [J]. Annals of neurology, 2019, 86(3): 427-35.

[4] ABRAHAM G, MALIK R, YONOVA-DOING E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke [J]. Nature communications, 2019, 10(1): 1-10.

[5] RITCHIE K. Mild cognitive impairment: an epidemiological perspective [J]. Dialogues in clinical neuroscience, 2022.

[6] LEONENKO G, BAKER E, STEVENSON-HOARE J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores [J]. Nat Commun, 2021, 12(1): 4506.

[7] VAN DEN BERG E, GEERLINGS M I, BIESSELS G J, et al. White matter hyperintensities and cognition in mild cognitive impairment and Alzheimer’s disease: a domain-specific meta-analysis [J]. Journal of Alzheimer's disease, 2018, 63(2): 515-27.

[8] ZACKOVá L, JáNI M, BRáZDIL M, et al. Cognitive impairment and depression: Meta-analysis of structural magnetic resonance imaging studies [J]. NeuroImage: Clinical, 2021, 32: 102830.

[9] REED E, NUNEZ S, KULP D, et al. A guide to genome-wide association analysis and post-analytic interrogation [J]. Statistics in medicine, 2015, 34(28): 3769-92.

[10] 牛晓歌. 基于大型前瞻性队列构建和评价中国人群脑卒中多基因遗传风险评分 [D]; 北京协和医学院, 2021.

[11] ANDREWS S J, FULTON-HOWARD B, GOATE A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease [J]. The Lancet Neurology, 2020, 19(4): 326-35.

[12] BELLOU E, BAKER E, LEONENKO G, et al. Age-dependent effect of APOE and polygenic component on Alzheimer's disease [J]. Neurobiology of aging, 2020, 93: 69-77.

[13] BONHAM L W, GEIER E G, FAN C C, et al. Age-dependent effects of APOE epsilon4 in preclinical Alzheimer's disease [J]. Ann Clin Transl Neurol, 2016, 3(9): 668-77.

[14] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System. KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016: 785-794 [Z]. 2016

[15] EATON J E, VESTERHUS M, MCCAULEY B M, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning [J]. Hepatology, 2020, 71(1): 214-24.

[16] TIBSHIRANI R. The lasso method for variable selection in the Cox model [J]. Statistics in medicine, 1997, 16(4): 385-95.

[17] LI J, LU Q, WEN Y. Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data [J]. Bioinformatics, 2020, 36(6): 1785-94.

[18] DUDBRIDGE F, GUSNANTO A. Estimation of significance thresholds for genomewide association scans [J]. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2008, 32(3): 227-34.

[19] EDMONDSON A C, BRAUND P S, STYLIANOU I M, et al. Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol [J]. Circulation: Cardiovascular Genetics, 2011, 4(2): 145-55.

[20] CHRISTODOULOU E, MA J, COLLINS G S, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models [J]. Journal of clinical epidemiology, 2019, 110: 12-22.

[21] RáCZ A, BAJUSZ D, HéBERGER K. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification [J]. Molecules, 2021, 26(4): 1111.

数据可用性声明：支撑本研究的科学数据已在中国科学院数据银行 ScienceDB 公开发布，访问地址为 http://cstr.cn/31253.11.sciencedb.j00150.00009，DOI：10.57760/sciencedb.j00150.00009，CSTR：31253.11.sciencedb.j00150.00009。